The Tech Against Terrorism Guidelines
–
Tech Company Transparency Reporting on Online Counterterrorism Efforts

Tech Against Terrorism – An initiative supported by UN CTED
contact@techagainstterrorism.org

The Tech Against Terrorism Guidelines

Tech company transparency reporting on online counterterrorism efforts

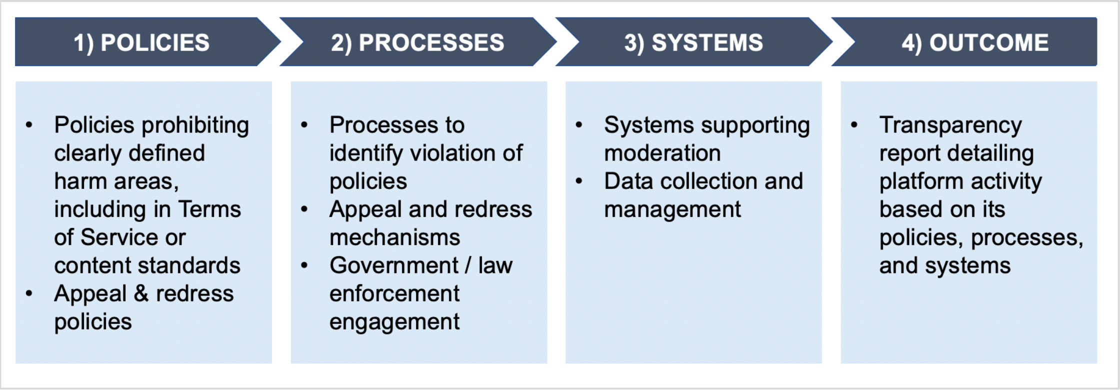
INTRODUCTION

Transparency is vital to ensure that the tech industry is accountable to the public and its users. Transparency reporting provides insight on to what extent fundamental freedoms such as freedom of expression and the right to privacy are respected across the internet. It can also encourage and recognise meaningful action from tech companies in tackling terrorist use of the internet and provide crucial insight on this threat. Transparency should therefore be considered a key aspect of counterterrorism online, and transparency reporting has been a core part of Tech Against Terrorism's support for the tech sector since 2017, including in the Mentorship Programme.

Larger tech platforms have made significant strides and now produce detailed transparency reports on their activities to remove content, however smaller tech platforms struggle to produce reports due to lack of resources and capacity. We regret that no government, to our knowledge, seems to have published meaningful transparency reports on their actioning requests and referrals to tech companies. To that end, we are launching the Tech Against Terrorism Guidelines for transparency reporting on online counterterrorism efforts to improve transparency and accountability across the tech and government sectors.

Introducing Tech Against Terrorism's Transparency Guidelines

Transparency is a process in which transparency reporting is an outcome. To enable reporting, companies need policies and moderation processes to support the generation of metrics that can be included in a report. Many smaller companies – where the majority of the terrorist threat is – might struggle to introduce such mechanisms. Even in cases where companies have these in place, they may not have the resources to build content management systems to manage data used for reporting. It is therefore not feasible to compel smaller tech companies to report on a large number of metrics and to hold them to the same standards as larger and longer-established platforms.



1) POLICIES
- Policies prohibiting clearly defined harm areas, including in Terms of Service or content standards
- Appeal & redress policies

2) PROCESSES
- Processes to identify violation of policies
- Appeal and redress mechanisms
- Government / law enforcement engagement

3) SYSTEMS
- Systems supporting moderation
- Data collection and management

4) OUTCOME
- Transparency report detailing platform activity based on its policies, processes, and systems

To ensure meaningful transparency across the tech industry, focus should be on a smaller number of core metrics to facilitate evaluation of company track records over time. Furthermore, there should be recognition of platform diversity. Mandated standardised reporting will not lead to meaningful transparency, as different platform purposes, policies, and processes will inevitably lead to discrepancies in data points.

The Tech Against Terrorism Guidelines
Tech company transparency reporting on online counterterrorism efforts

These Guidelines complement the practical work Tech Against Terrorism has carried out in support of the tech industry since 2017, particularly through the Mentorship Programme. This programme focusses on improving content standards, content moderation practices, and transparency reporting, and we encourage all mentee companies to be transparent about their policies and practices. It is our aim that the Guidelines can provide an improved framework to support these activities.

The Guidelines focus on encouraging more companies to all report on (at least) a core set of metrics that reflect the holistic process outlined above. The Guidelines are meant to allow enough flexibility to encourage transparency from all types of tech companies on all types of terrorist activity. By publishing these Guidelines, we hope to set a realistic target for smaller tech companies and to galvanise support for transparency, that is both meaningful and realistic for the companies that are most often exploited by terrorist groups, and have tangible impact in improving transparency

TECH AGAINST TERRORISM'S TRANSPARENCY REPORTING GUIDELINES

In Part A we ask platforms to detail some of their core policies. In Part B we ask companies to provide detail on their moderation processes and systems. In Part C we ask companies to report on quantitative metrics to understand trends and patterns with regard to terrorist use of their platform.

Part A: Policies

Describe the moderation policies you have introduced in your Terms of Service, Community Guidelines, and/or content standards to tackle terrorist use of your platform by detailing your platform's:

1. Working definition and/or prohibition of terrorism[1]
2. Appeal and redress mechanism[2]

Part B: Moderation processes and systems

Describe your moderation processes and systems by detailing the following areas:

Discovery

Your platform's:

3. Processes for detecting terrorist content and/or activity
4. Systems and tools used to detect terrorist content and/or activity
5. Processes and systems to facilitate external reporting and flagging of terrorist content and/or activity (if applicable)

[1] This can either the platform's own definition, a definition created by government body or civil society group or expert researchers, or a reference to a designation list.

[2] Here, we encourage platforms to base their appeals process on the threshold outlined in the Santa Clara Principles: https://santaclaraprinciples.org/

Enforcement

Your platform's:

6. Processes for actioning detected terrorist content and/or activity
7. Systems and tools for actioning detected terrorist content and/or activity

Data collection

Your platform's:

8. Processes *and/or* systems used to collect moderation statistics

Due diligence:

9. Describe what due diligence and/or verification your platform conducts with regard to the processes and systems you use

Part C: Moderation statistics

In the quantitative part of the report, we ask platforms to detail:

10. Discovery of terrorist content and/or activity, in total numbers and segmented by:
    a. Government and law enforcement requests
        i. Broken down by country of origin
    b. Government and law enforcement ToS referrals[3]
    c. Proactive discovery via the processes and/or systems outlined in Part B
    d. User reports
    e. GIFCT hash-sharing database (if relevant)

11. Actioning of terrorist content and/or activity, in total numbers or percentages and segmented to showcase where actioning was the result of:
    a. Government and law enforcement requests
        i. Broken down by country of origin
    b. Government and law enforcement ToS referrals
    c. Proactive discovery
    d. User reports
    e. GIFCT hash-sharing database (if relevant)

12. User appeals *and* appeal success rate

[3] Unfortunately, it is not always evident from where government ToS referrals emanate or whether it is in fact a government that has referred the content. We therefore encourage governments to improve transparency about such referrals.

Example report based on the Guidelines

The below is an example report based on what reporting in accordance with the Guidelines could look like. The Guidelines provide a baseline for reporting, and it is up to each company to decide what depth they wish to provide on each point.

A: Policies	B: Processes & Systems
• *"At Platform terrorist activity is prohibited. By **terrorist activity we mean activity, included content-sharing, carried out to benefit terrorist groups listed in the UN Security Council Consolidated List**"* • *"If you believe your content has been misidentified and wrongly removed as terrorism, **you can appeal contacting us as at appeals@platform.com – we aim to review all appeals within 2 weeks** and will notify you of our decision"*	• *At Platform we have a dedicated **team working to detect terrorist content, but we also rely on user reports and the Terrorist Content Analytics Platform** to identify terrorist activity. We also work with Tech Against Terrorism as part of a trusted flagging programme. Once we detect suspected terrorist activity we **review it manually within our Trust & Safety team before making a decision** on whether to remove it or limiting its spread"* • *"We use **image hashing to detect suspected terrorist content**, but we **always review content manually**"* • *"We report and **record all activity on our moderation statistics in a dedicated database**. This work is done in accordance with GDPR"*

C: Moderation statistics					
			2019	**2020**	**2021**
Terrorist content* referral or discovery					
	Government removal requests		250	350	450
		Breakdown by country			
	Government ToS referrals		150	250	350
		Breakdown by country (optional)			
	User referral		15	26	57
	Proactive discovery		400	400	400
	GIFCT hash-sharing database		100	150	200
Terrorist content removal or actioning %					
	Government removal requests		80%	90%	95%
		Breakdown by country			
	Government ToS referrals		80%	90%	95%
		Breakdown by country (optional)			
	User referral		40%	40%	40%
	Proactive discovery		95%	95%	95%
	GIFCT hash-sharing database		90%	90%	90%
Appeals					

	Total		2	5	10
	% successful		90%	90%	90%

Glossary of key terms	
“Working definition and/or prohibition of terrorism”	This can be a clear reference to a designation list or to a government and/or academic definition, or to the definition that your platform has opted to develop yourselves (if relevant).
“Appeal”	For the purposes of the Guidelines, appeals refer to a process in which users can ask companies to address perceived errors on the company’s part with regards to actions taken to remove and/or disrupt content, activity, and accounts discovered under the company’s counterterrorism efforts.
“Redress”	For the purposes of the Guidelines, redress refers to actions taken by tech companies to address errors flagged by users via the appeal process, for example content and/or account reinstatement.
“Processes”	Any process or workflow that your platform has introduced that supports discovery, moderation, and/or statistics collection of terrorist content and/or activity, including (but not limited to) flagger programmes, staffing, and usage of voluntary cooperative frameworks such as the TCAP.
“Systems”	Any systems or tooling, such as automated data-driven tools, that your platform has introduced that supports discovery, moderation, and/or statistics collection of terrorist content and/or activity.
“Actioning”	For the purposes of this document, “actioning” means any moderation enforcement decision made by a company, and may include measures such as (but not limited to) content removal, account removal, temporary suspensions, restricted access, and disabling of access.
“Government and law enforcement request”	Request to moderate content activity submitted by a government-affiliated body or law enforcement agency via appropriate legal channels, including court orders or other clearly legally defined channels, that references content illegality under specified legal framework.
“Government and law enforcement ToS referral”	Flagging of content / activity made by a government-affiliated body or law enforcement agency, sometimes via extra-legal channels, for companies to examine against their own policies and content standards.
“Proactive discovery”	Activities you as a company undertake on your own initiative to discover and surface terrorist content or activity on your platforms. This includes but is not limited to the use automated tooling.
“User reports”	Reports of suspected content or activity made by users of the platform. This may also include trusted flagging schemes.

TRANSPARENCY REPORTING: KEY CONSIDERATIONS

1. Smaller platforms should be prioritised

The majority of terrorist and violent extremist activity occurs on smaller platforms. Terrorists exploit such platforms due to their lack of capacity to respond to the threat, which is why support mechanisms for smaller companies are needed. If the aim is to increase tech sector transparency on terrorist use of the internet and how the tech sector responds, we need to ensure that the platforms most used by terrorists are able to produce reports.

2. Smaller platforms may lack capacity

Smaller platforms often do not have resources to implement or maintain processes that support transparency reporting across a multitude of metrics. Given that a majority of terrorist activity online takes place on smaller platforms, it is important to encourage improvement by setting realistic targets for smaller companies, and to where possible give smaller platforms leniency if they due to lack of capacity fail to comply or make mistakes.

3. Pay attention to process

Transparency reporting is an outcome that starts with introducing policies and moderation processes. Many smaller companies might not have these in place. It is important to not compel the smallest platforms to produce reports with a large amount of detailed metrics for which they do not have the policies, processes, or tools to develop. Instead, focus should be on supporting platforms implement such processes.[4]

4. Diversity

Different platforms will have different purposes, policies, and processes in place. This should be encouraged as part of a diverse and vibrant internet. In the absence of international consensus around key definitions of terrorism and/or violent extremism, companies use varying working definitions of these terms and will therefore remove different types of content under such policies. A platform aimed at children might not have a specific policy on terrorism, but instead include it as part of a wider “abusive / violent content” category. Even platforms that do explicitly prohibit terrorism might differ in what content or activity they prohibit, as some platforms might allow terrorist content if shared for educational or journalistic purposes, whereas others will not. Furthermore, business models are likely to impact the resources and tooling platforms have in place. For example, a discussion platform might be better equipped to detect hate speech or misinformation in text, whereas a visually driven platform might be better equipped to tackle high-priority harm areas such as child sexual abuse imagery or copyright infringement. Such diversity means that mandating standardised reporting will make companies squeeze data points into pre-made categories which will not render transparency reporting meaningful.

5. Rule of law and incentives

[4] Support around policy, moderation, and transparency reporting are key parts of the Tech Against Terrorism Mentorship and Membership programmes: https://www.techagainstterrorism.org/membership/tech-against-terrorism-mentorship/

Transparency reporting should not be used as a tool by which to introduce new content moderation demands for tech platforms in an extra-legal fashion. The risk with compelling companies to report on a pre-determined number of metrics without consideration to platform capacity, policy, or process is that we create incentives for companies to remove content for which there is no legal basis, or that platforms will over-zealously remove content to placate policy-makers. Such risks are likely exacerbated by the fact that there is little consensus around key terms like terrorism and/or terrorist content. Any transparency reporting demands therefore need to be underpinned by a clear legal basis, for example via designation of terrorist groups.

6. Meaningfulness

Whilst transparency is a goal in itself, we should strive towards encouraging transparency that leads to meaningful insight. To ensure that such meaningful transparency is practical for smaller companies, we suggest focussing on a smaller number of transparency metrics to facilitate evaluation of platforms' records over time.

7. Reciprocity

We also see transparency reporting as an important way to encourage governments to uphold human rights and fundamental freedoms. We encourage governments to publish transparency reports regarding their content referrals and removal requests.

Further reading

- Tech Against Terrorism Pledge
- Tech Against Terrorism Mentorship Programme
- Santa Clara Principles on Transparency and Accountability in Content Moderation
- Open Technology Institute's Transparency Reporting Toolkit
- EFF Deep Links: "Thank You For Your Transparency Report, Here's Everything That's Missing"
- Daphne Keller, Center for Internet Society, Stanford University: "Some Humility About Transparency"

ANNEX

Annex 1. Tech Against Terrorism Pledge

Introduction

The increased exploitation of information and communication technologies for terrorist and violent extremist purposes raises new challenges related to countering terrorism whilst respecting human rights, in particular with regards to freedom of expression and privacy. In the context of preventing and countering terrorism and violent extremism, effective counter-measures and the protection of human rights are not conflicting goals, but complementary and mutually reinforcing.

Tech Against Terrorism has developed six guiding principles (the Tech Against Terrorism Pledge) which inform our approach and underpin our framework for engaging with the very smallest technology companies.[1] These reinforce the importance of addressing challenging content and will support small tech companies in articulating their commitment to human rights and diversity in transparent, accountable, and collaborative ways. Our pledge complements the Global Network Initiative (GNI)[2] Principles as it is specifically designed for smaller tech platforms.

The Tech Against Terrorism Pledge provides simple and accessible guidelines to help the very smallest companies understand the importance of tackling terrorist exploitation in a manner that respects human rights and freedom of speech. With our pledge, we want to ensure that small companies – who often do not have enough resources to familiarise themselves with the myriad of legal regimes and social contexts which may apply to their services – can contribute to a free internet. The pledge is a starting point from which companies can build their own appropriate systems and policies. Company commitments to the pledge should be understood as aspirations to be achieved as quickly and thoroughly as possible, consistent with available resources and scale.

Our pledge is based on the GNI Principles and internationally recognised norms as articulated in the Universal Declaration of Human Rights ("UDHR"), the International Covenant on Civil and Political Rights ("ICCPR"), the International Covenant on Economic, Social and Cultural Rights ("ICESCR"), UN Security Council resolutions and documents S/RES/1624 (2005), S/RES/2129 (2013), S/RES/2322 (2016), S/RES/2354 (2017) and S/2017/375, and the UN Guiding Principles on Business and Human Rights ("UN Guiding Principles"). These constitute crucial normative precepts to help technology companies tackle exploitation of their services in a manner that promotes and protects human rights.[3]

The Tech Against Terrorism Pledge

1. **Freedom of Expression**

"We respect the right to freedom of expression that should be enjoyed by our users and will take actions consistent with applicable law to protect it from unlawful or unnecessary restrictions."

Article 19 of the ICCPR provides that "*1. Everyone shall have the right to hold opinions without interference. 2. Everyone shall have the right to freedom of expression; this right shall include* *freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.*

The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals."

2. **Non-Discrimination and Diversity**

"We respect the right of our users to express diverse views and opinions, and commit to educating users regarding what content and expression is not permitted on our platforms through clear terms of service and their transparent and consistent application."

> Article 24 of the ICCPR states that *"All persons are equal before the law and are entitled without any discrimination to the equal protection of the law."* Article 15 of the ICESCR recognises the rights of everyone to take part in cultural life.

3. **Privacy**

"We respect the privacy of all our users and will take actions consistent with applicable law to protect it from arbitrary or unlawful interference."

> UNDHR Article 12 and ICCPR Article 17 states *"No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks."*

4. **Transparency and Accountability**

"We appreciate the need to account for what content we deem impermissible on our platforms, how we address government requests related to content on our platforms, and how we make determinations about content. To this end, we value and strive for transparency regarding those policies and practices, especially with regard to how they may impact the above-mentioned human rights-principles."

> Guiding Principle 21 articulates an expectation that companies will account for how they address human rights and the commentary further explains that this "*requires that business enterprises have in place policies and processes through which they can both know and show that they respect human rights in practice. Showing involves communication, providing a measure of transparency and accountability to individuals or groups who may be impacted and to other relevant stakeholders, including investors.*"

5. **Remedy**

"While we strive to apply content policies fairly and consistently, we recognise that resource limitations, cultural contexts, and other factors may result in decisions that unintentionally cause negative impacts. To address this eventuality, we commit to devising appropriate mechanisms to allow individuals impacted by our policies and practices to bring information to our attention."

Guiding Principle 20 states: *"To make it possible for grievances to be addressed early and remediated directly, business enterprises should establish or participate in effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted."*

6. Collaboration

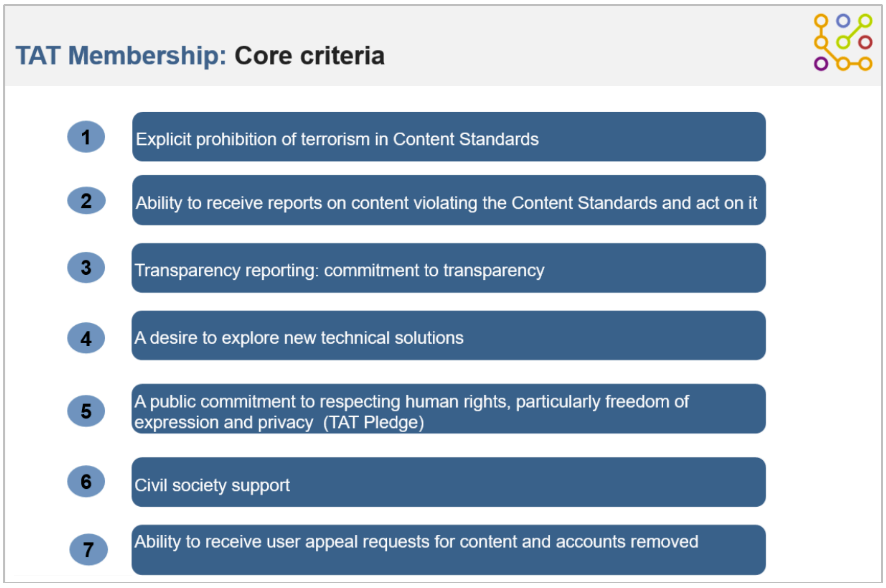
"We commit to work with partner organisations and enterprises to collaboratively develop strategies to keep our platforms and products safe from abuse by terrorist organisations and their supporters, and to promote tolerance, coexistence and diversity."

Article 19 of the ICCPR states that the exercise of freedom of expression carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order of public health or morals.

S/RES/1624 (2005) calls upon States to prohibit by law incitement to commit a terrorist act and S/RES/2354 (2017) condemns "*in the strongest terms the incitement of terrorist acts" and repudiates "attempts at the justification or glorification of terrorist acts that may incite further terrorist acts."*

S/RES/2354 (2017) further stresses the importance of the role of the business community *"in efforts to enhance dialogue and broaden understanding, and in promoting tolerance and coexistence, and in fostering an environment which is not conducive to incitement of terrorism, as well as in countering terrorist narratives."* It urges further development of initiatives to strengthen public-private partnerships in this area, and notes the benefits of engagement with a wide range of actors, including youth, families, women, community leaders, and other concerned groups of civil society.

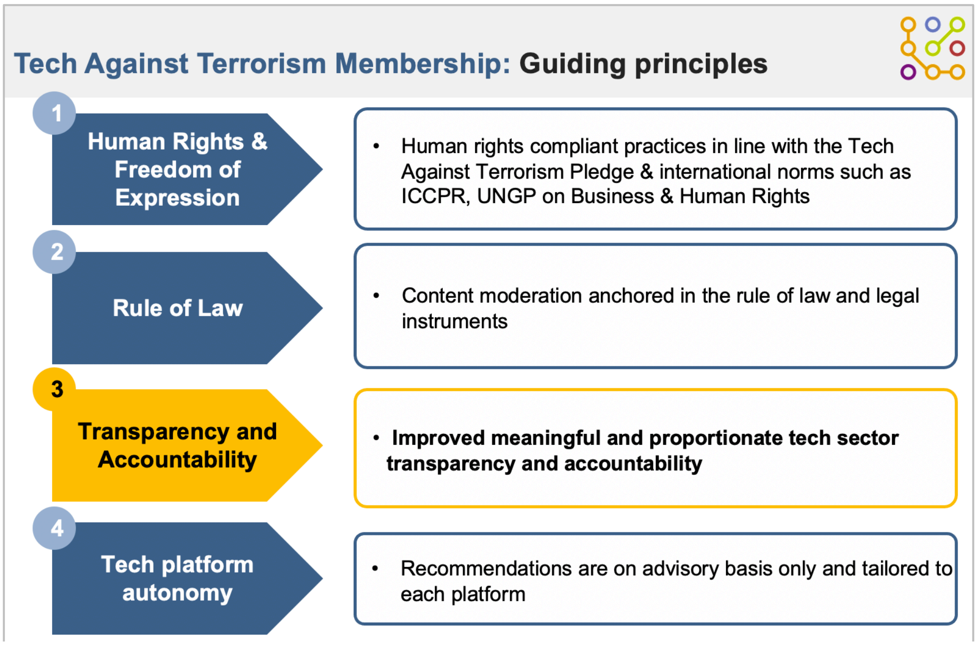
Annex 2. Tech Against Terrorism Membership Criteria



TAT Membership: Core criteria

1. Explicit prohibition of terrorism in Content Standards
2. Ability to receive reports on content violating the Content Standards and act on it
3. Transparency reporting: commitment to transparency
4. A desire to explore new technical solutions
5. A public commitment to respecting human rights, particularly freedom of expression and privacy (TAT Pledge)
6. Civil society support
7. Ability to receive user appeal requests for content and accounts removed

Annex 3. Tech Against Terrorism Membership: Core Principles



Tech Against Terrorism Membership: Guiding principles

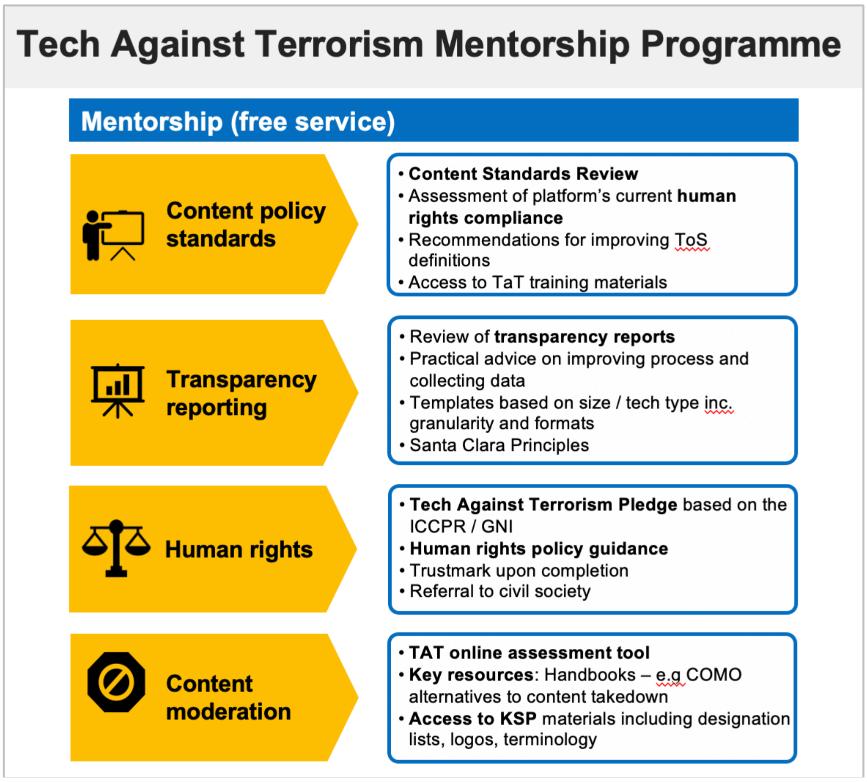
1 **Human Rights & Freedom of Expression**
- Human rights compliant practices in line with the Tech Against Terrorism Pledge & international norms such as ICCPR, UNGP on Business & Human Rights

2 **Rule of Law**
- Content moderation anchored in the rule of law and legal instruments

3 **Transparency and Accountability**
- **Improved meaningful and proportionate tech sector transparency and accountability**

4 **Tech platform autonomy**
- Recommendations are on advisory basis only and tailored to each platform

Annex 4. Summary of Tech Against Terrorism Mentorship Programme



Tech Against Terrorism Mentorship Programme

Mentorship (free service)

Content policy standards
- **Content Standards Review**
- Assessment of platform's current **human rights compliance**
- Recommendations for improving ToS definitions
- Access to TaT training materials

Transparency reporting
- Review of **transparency reports**
- Practical advice on improving process and collecting data
- Templates based on size / tech type inc. granularity and formats
- Santa Clara Principles

Human rights
- **Tech Against Terrorism Pledge** based on the ICCPR / GNI
- **Human rights policy guidance**
- Trustmark upon completion
- Referral to civil society

Content moderation
- **TAT online assessment tool**
- **Key resources**: Handbooks – e.g COMO alternatives to content takedown
- **Access to KSP** materials including designation lists, logos, terminology